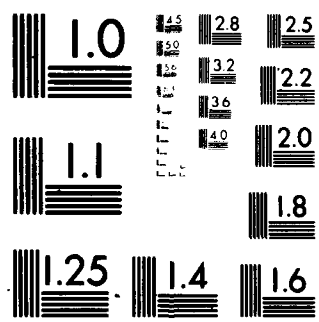
AD-A088 319 MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR COMPUTE--ETC F/G 9/2
COMPLEXITY OF STORAGE AND RETRIEVAL PROBLEMS.(U)
JUL 80 P ELIAS DAAG29-77-C-0012
UNCLASSIFIED ARO-14649.2-EL NL

1 of 1
AD A
088319

END
DATE
FILMED
9-80
DTIC



MICROCOPY RESOLUTION TEST CHART

ARO 14649.2-EL

(12)

B.S.

AD A088319

TITLE

COMPLEXITY OF STORAGE AND RETRIEVAL PROBLEMS

TYPE OF REPORT (TECHNICAL, FINAL, ETC.)

FINAL REPORT

AUTHOR (S)

PETER ELIAS

DATE

18 JULY 1980

U. S. ARMY RESEARCH OFFICE

CONTRACT / GRANT NUMBER

DAAG29-77-C-0012

INSTITUTION

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER / 2. GOVT ACCESSION NO. AD-A088 319 (9)	3. ~~RECIPIENT'S CATALOG NUMBER~~ Final rept.
4. TITLE (and Subtitle) (6) Complexity of Storage and Retrieval Problems.	5. ~~TYPE OF~~ REPORT & PERIOD COVERED FINAL; 14 Jan 1977 ~~to~~ 30 May 1980
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) Peter Elias	8. CONTRACT OR GRANT NUMBER(s) (15) DAAG29-77-C-0012
9. PERFORMING ORGANIZATION NAME AND ADDRESS Laboratory for Computer Science Massachusetts Institute of Technology	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709 (11)	12. ~~REPORT DATE~~ 18 July 1980
	13. ~~NUMBER OF PAGES~~
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (12) 6 (18) ARO (17) 14647.2-EL	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE NA

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

information, storage, retrieval, complexity, algorithms, coding, universal codes.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This project has investigated the costs imposed by the need for communication between user and machine and between components of the machine in solving data-processing problems. There have been three principal subtasks. The first, exploration of the minimal costs of storing and accessing information in simple data structures, is the oldest and results have been reported in several publications. The second is the design of minimax optimal universal codeword sets which can be used to represent any message set efficiently by assigning messages

DD FORM 1473 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

499 648

20. Abstract (continued from previous page)

in order of decreasing probability to codewords in order of increasing length. A manuscript giving results of this research is almost ready for publication. The third is the exploration of switching networks which can be used for communication between a number of processors engaged in a common computational task: that work is still in an early stage and is not yet ready for publication.

Dist. | Special
A

BODY

When computers carry out interactive data processing tasks, communication is necessarily involved. Commands must be communicated from the user to the processor, results must be communicated from the processor to the user, and in the course of carrying out the computation there must be communication between the processor and the memory hierarchy. If multiple processors are involved in carrying out portions of a common task then they must communicate with one another. The basic objective of the research carried out in this project has been to explore the contribution of these communications requirements to the difficulty of carrying out data-processing tasks.

The storage and retrieval of information presents a class of problems for which such an investigation is especially appropriate, since the costs of a storage and retrieval problem are largely the costs of communication between processor and memory and the costs of storing information in memory. Information theory provides results and techniques which are appropriate for investigating such problems, and the principal published work resulting from this research has been an analysis of the trading relations between storage and access costs for best-possible representations of simple data structures. A review paper by Elias [1], reporting work performed under a previous ARO contract, gives results in special cases. The doctoral thesis of Donna Brown, issued in [2] as a report of the Laboratory for Computer Science, explores trading relations for linear data structures- lists, stacks and queues- in detail. Those and related results have been published by Brown in [3] and [4], with acknowledgement of ARO support under this contract, and another publication by Brown is in preparation. So is a publication by Elias, which generalizes the results in [1].

One obvious application of information theory to storage and retrieval problems concerns the choice of representation for data values. An entry in a data base which is selected from a fixed set can be assigned a codeword which is a sequence of symbols, and if the set of possible values is large and values occur with different frequencies then Huffman encoding can be used to reduce the average amount of storage space required by using the shortest codewords for the most frequent values. A difficulty with this approach is that in data processing the exact knowledge of frequencies of different values which is assumed in information theory and is available for English text may not be available for a given application. A second difficulty is that a different codebook must be consulted in looking up each value selected from a different set. Both of these problems are alleviated by using a universal codeword set, as introduced in [5]. A universal set of codewords is defined to have the property that if messages in a message set are assigned in order of decreasing probability to codewords in order of increasing length, the ratio of the

resulting average codeword length to the length of a best possible (Huffman) encoding of that source is bounded by a constant for all message sets whose entropy is neither zero nor infinite. In research reported in [5] and supported by a predecessor ARO contract I copnstructed infinite universal codeword sets and showed that the ratio of the average codeword length for such a set to the average length of a Huffman code was bounded by 3 for the worst possible probability distribution. However that bound was not shown to be the actual value of the ratio in the worst case (which turns out to be 2, not 3) nor was that set of codewords shown to minimize the maximum of that ratio over all message sets. More recent research by Rissanen [6] and by Davisson and Leon-Garcia [7] used different measures of performance than that ratio, and restricted their attention to finite rather than infinite codeword sets, but were able to find codeword sets which were minimax optimal by their measures on the sets of message probability distributions they considered. In recent work [8] now being prepared for publication I have found fast algorithms for the design of minimax optimal universal codeword sets by the ratio cost measure: the average codeword length for such a code is at most 253/160 ~ 1.58 times the average codeword length for a Huffman code for the worst message distribution.

One other topic has also been explored, by a graduate student Andrew Boughton who has been supported in part under this contract. Boughton has investigated the communications problems involved in using a large number of processors in parallel to solve a single problem. The obvious way of making it possible to connect the output of one processor to the input to any other is a crossbar, which takes a number of elements growing like the square of n to connect n processors, which becomes expensive for large n. Other connection network techniques are informationally more efficient and require only nlogn elements to connect n devices. However such networks may require long wires, and thus both significantly greater delay and significantly greater implementation cost when built as integrated circuits, in which wire is as expensive as devices. This work is also not yet ready for publication, but Boughton's doctoral thesis proposal will be complete shortly.

BIBLIOGRAPHY

1) Elias, P., "An information-theoretic approach to computational complexity", in Topics in Information Theory, North-Holland Publishing Co., N.Y., (1977) pp.171 to 198.

2) Brown, D. J., "Storage and access costs for implementations of variable-length lists", Tech. Report MIT/LCS/TR-217, MIT, Cambridge, Ma. (1979).

3) Brown, D. J., "Kraft storage and access for list implementations", pp.100-106, Proc. 12th annual ACM Symposium on Theory of Computing, ACM (1980).

4) Brown, D. J., "Information-theoretic bounds for implementing stacks", to appear in Proc. 14th annual Conference on Information Sciences and Systems (1980).

5) Elias, P., "Universal codeword sets and representations of the integers", pp. 194-203, IEEE Transactions on Information Theory 21 (March 1975).

6) Rissanen, J., "Minimax codes for finite alphabets", pp. 389-392, IEEE Transactions on Information Theory 24 (May 1978).

7) Davisson, L. D. and Leon-Garcia, A. "A source matching approach to finding minimax codes", pp.166-174, IEEE Transactions on Information Theory 26 (March 1980).

PARTICIPATING SCIENTIFIC PERSONNEL

Peter Elias, principal investigator.

Donna J. Brown, Research Assistant. Received Ph. D. in Dept. of Electrical Engineering and Computer Science September 1978.

Andrew J. Boughton, Research Assistant and current doctoral candidate.

Andrew Moulton, Research Assistant and current doctoral candidate.